Mellanox CloudX, Mirantis Fuel 5.1/ 5.1.1 Solution Guide

Rev 1.1

www.mellanox.com

NOTE:
THIS HARDWARE, SOFTWARE OR TEST SUITE PRODUCT (“PRODUCT(S)”) AND ITS RELATED DOCUMENTATION ARE PROVIDED BY MELLANOX TECHNOLOGIES “AS-IS” WITH ALL FAULTS OF ANY KIND AND SOLELY FOR THE PURPOSE OF AIDING THE CUSTOMER IN TESTING APPLICATIONS THAT USE THE PRODUCTS IN DESIGNATED SOLUTIONS. THE CUSTOMER'S MANUFACTURING TEST ENVIRONMENT HAS NOT MET THE STANDARDS SET BY MELLANOX TECHNOLOGIES TO FULLY QUALIFY THE PRODUCTO(S) AND/OR THE SYSTEM USING IT. THEREFORE, MELLANOX TECHNOLOGIES CANNOT AND DOES NOT GUARANTEE OR WARRANT THAT THE PRODUCTS WILL OPERATE WITH THE HIGHEST QUALITY. ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT ARE DISCLAIMED. IN NO EVENT SHALL MELLANOX BE LIABLE TO CUSTOMER OR ANY THIRD PARTIES FOR ANY DIRECT, INDIRECT, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES OF ANY KIND (INCLUDING, BUT NOT LIMITED TO, PAYMENT FOR PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY FROM THE USE OF THE PRODUCT(S) AND RELATED DOCUMENTATION EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Mellanox Technologies
350 Oakmead Parkway Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

Mellanox Technologies, Ltd.
Beit Mellanox
PO Box 586 Yokneam 20692
Israel
www.mellanox.com
Tel: +972 (0)74 723 7200
Fax: +972 (0)4 959 3245

© Copyright 2014. Mellanox Technologies. All Rights Reserved.

Mellanox®, Mellanox logo, BridgeX®, ConnectX®, Connect-IB®, CoolBox®, CORE-Direct®, InfiniBridge®, InfiniHost®, InfiniScale®, MetroX®, MLNX-OS®, PhyX®, ScalableHPC®, SwitchX®, UFM®, Virtual Protocol Interconnect® and Voltaire® are registered trademarks of Mellanox Technologies, Ltd.

ExtendX™, FabricIT™, Mellanox Open Ethernet™, Mellanox Virtual Modular Switch™, MetroDX™, TestX™, Unbreakable-Link™ are trademarks of Mellanox Technologies, Ltd.

All other trademarks are property of their respective owners.

Table of Contents

Preface **4**
1 Overview **5**
2 Virtualization **6**
2.1 eSwitch Capabilities and Characteristics 7
2.2 Performance Measurements 7
3 Storage Acceleration **8**
4 Networking **10**
4.1 Network Types 11
4.1.1 Admin (PXE) Network 11
4.1.2 Storage Network 11
4.1.3 Management Network 11
4.1.4 Private Networks 11
4.1.5 Public Network 12
4.2 Physical Connectivity 14
4.3 Network Separation 14
4.4 Lossless Fabric (Flow-Control) 14
5 Requirements **15**
5.1 Hardware Requirements 15
5.2 Operating Systems 16
6 Rack Configuration **17**
6.1 32 Compute Nodes Setup with HA (3 Controller Nodes) 17
7 Installation and Configuration **18**

Preface

About this Manual

This manual is a reference architecture and an installation guide for a small size OpenStack cloud of 2-32 compute nodes based on Mellanox interconnect hardware and Mirantis Fuel software.

Audience

This manual is intended for IT engineers, system architects and anyone interested in understanding or deploying Mellanox CloudX™ using Mirantis Fuel.

Related Documentation

For additional information, see the following documents:

Document	Location
Mellanox OpenStack Reference Architecture	http://www.mellanox.com/openstack/
Mellanox MLNX-OS User Manual	http://support.mellanox.com/ NOTE: active support account required to access manual.
HowTo Install Mirantis Fuel 5.1/5.1.1 OpenStack with Mellanox Adapters Support	http://community.mellanox.com/docs/DOC-1474
HowTo Configure 56GbE Link on Mellanox Adapters and Switches	http://community.mellanox.com/docs/DOC-1460
Firmware - Driver Compatibility Matrix	http://www.mellanox.com/page/mlnx_ofed_matrix?mtag=linux_sw_drivers
Mirantis Openstack installation guide	http://docs.mirantis.com/openstack/fuel/fuel-5.1/

Revision History

Revision	Description
1.0	First document
1.1	Minor changes for 5.1.1 support.

1 Overview

Mellanox CloudX™ is reference architecture for the most efficient cloud infrastructure which makes use of open source cloud software (i.e. OpenStack) while running on Mellanox® interconnect technology. CloudX utilizes off-the-shelf building blocks (servers, storage, interconnect and software) to form flexible and cost-effective private, public, and hybrid clouds. In addition, it incorporates virtualization with high-bandwidth and low-latency interconnect solutions while significantly reducing data center costs. Built around the fastest interconnect technology of 40Gb/s and 56Gb/s Ethernet, CloudX provides the fastest data transfer and most effective utilization of computing, storage and Flash SSD components.

Based on Mellanox high-speed, low-latency converged fabric, CloudX provides significant cost reductions in CAPEX and OPEX through the following means:

- High VM rate per compute node
- Efficient CPU utilization due to hardware offloads
- High throughput per server for compute and hypervisor tasks
- Fast, low-latency access to storage

Mirantis OpenStack is one of the most progressive, flexible, open distributions of OpenStack. In a single commercially supported package, Mirantis OpenStack combines the latest innovations from the open source community with the testing and reliability expected of enterprise software.

The integration of Mirantis Fuel software and Mellanox hardware generates a solid and high performing solution for cloud providers.

The solution discussed in this guide is based on Mirantis OpenStack 5.1/5.1.1 (OpenStack Icehouse release).

Single root I/O virtualization (SR-IOV) based networking and iSER block storage over the Mellanox ConnectX-3 adapter family are integrated into Mirantis OpenStack which offer the following features:

- Fabric speeds of up to 56GbE based on Mellanox SX1036 Ethernet switch systems
- iSER (High performance iSCSI protocol over RDMA) storage transport for Cinder
- SR-IOV high performance VM links by Mellanox SR-IOV plugin for OpenStack (included in the ML2 plugin)

2 Virtualization

SR-IOV allows a single physical PCIe device to present itself as multiple devices on the PCIe bus. Mellanox ConnectX®-3 adapters are capable of exposing up to 127 virtual instances called virtual functions (VFs). These VFs can then be provisioned separately. Each VF can be viewed as an additional device associated with a physical function (PF). The VF shares the same resources with the PF, and its number of ports equals those of the PF.

SR-IOV is commonly used in conjunction with an SR-IOV enabled hypervisor to provide virtual machines (VMs) with direct hardware access to network resources, thereby improving performance.

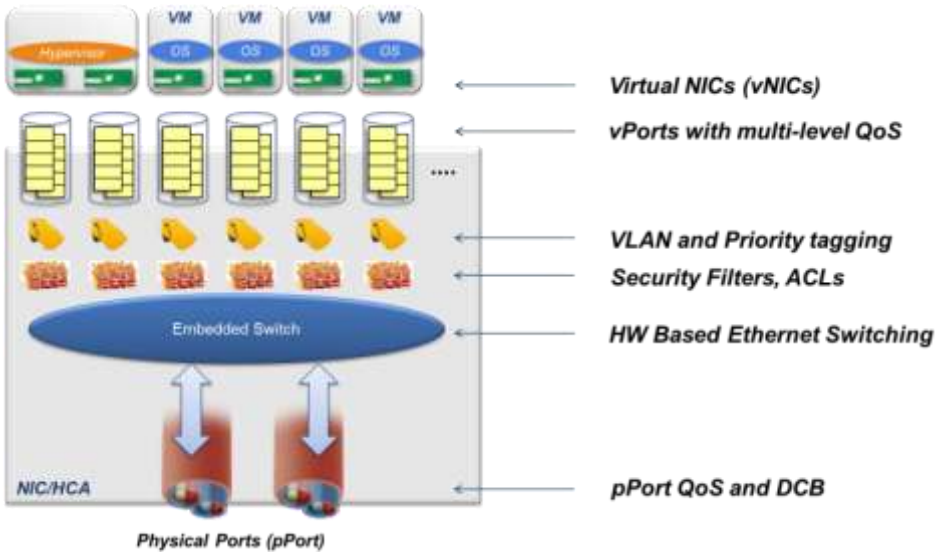
Mellanox ConnectX-3 adapters equipped with an onboard embedded switch (eSwitch) are capable of performing Layer-2 (L2) switching for the different VMs running on the server. Using the eSwitch yields even higher performance levels, and improves security and isolation.

The installation is capable of handling Mellanox NIC cards. It updates to the proper firmware version which incorporates SR-IOV enablement and defines 16 VFs by default. Each spawned VM is provisioned with one VF per network attached. The solution supports up to 16 VMs on a single compute node connected to single network or 8 VMs connected to 2 networks or any other combination which sums up to 16 networks in total. To support more than 16 vNICs, contact Mellanox Support.

Note: SR-IOV support for OpenStack is under development. Security groups are not supported with SR-IOV.

If the setup is based on Mellanox OEM NICs, make sure to have a firmware version compatible with OFED version 2.2-1.0.0 (FW version 2.31.5050) or later. Make sure that this firmware version supports SR-IOV. Consult vendor for more information.

Figure 1 - eSwitch Architecture



2.1 eSwitch Capabilities and Characteristics

The main capabilities and characteristics of eSwitch are as follows:

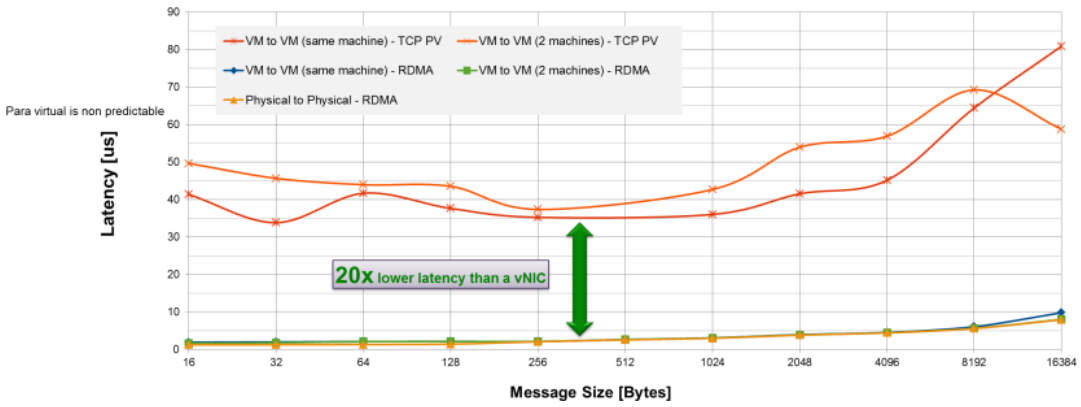
- Virtual switching: Creating multiple logical virtualized networks. The eSwitch offload engines handle all networking operations to the VM, thereby dramatically reducing software overheads and costs.
- Performance: Switching is handled by hardware as opposed to other applications that use a software-based switch. This enhances performance by reducing CPU overhead.
- Security: The eSwitch enables network isolation (using VLANs) and anti-MAC spoofing.
- Monitoring: Port counters are supported.

2.2 Performance Measurements

Many data center applications benefit from low-latency network communication while others require deterministic latency. Using regular TCP connectivity between VMs can create high latency and unpredictable delay behavior. Figure 2 exhibits the dramatic difference (20X improvement) delivered by SR-IOV connectivity running RDMA compared to para-virtualized vNIC running a TCP stream.

Using the direct connection of SR-IOV and ConnectX-3, the hardware eliminates software processing which delays packet movement. This results in consistent low-latency by allowing application software to rely on deterministic packet-transfer times.

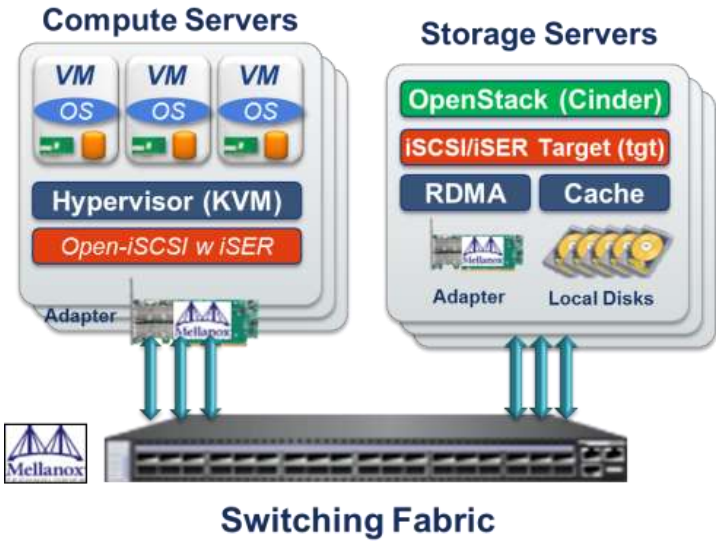
Figure 2 - Latency Comparison



3 Storage Acceleration

Data centers rely on communication between compute and storage nodes as compute servers read and write data from storage servers constantly. To maximize the server's application performance, communication between the compute and storage nodes must have the lowest possible latency and CPU utilization, and the highest possible bandwidth.

Figure 3 - OpenStack Based IaaS Cloud POD Deployment Example



Storage applications, relying on iSCSI over TCP communications protocol stack, continuously interrupt the processor to perform basic data movement tasks (packet sequence and reliability tests, reordering, acknowledgements, block level translations, memory buffer copying, etc). This causes data center applications that rely heavily on storage communication to suffer from reduced CPU efficiency as the processor is busy sending data to and from the storage servers rather than performing application processing. The data path for applications and system processes must wait in line with protocols such as TCP, UDP, NFS, and iSCSI for their turn to use the CPU. This not only slows down the network, but also uses system resources that could otherwise have been used for executing applications faster.

Mellanox OpenStack solution extends the Cinder project by adding iSCSI running over RDMA (iSER). By leveraging RDMA, Mellanox OpenStack delivers 6X better data throughput (for example, increasing from 1GB/s to 6GB/s) while simultaneously reducing CPU utilization by up to 80% (see Figure 4).

Mellanox ConnectX®-3 adapters bypass the operating system and CPU by using RDMA, thereby allowing much more efficient data movement. iSER capabilities are used to accelerate hypervisor traffic, including storage access, VM migration, and data and VM replication. The use of RDMA shifts data movement processing to the Mellanox ConnectX-3 hardware, which provides zero-copy message transfers for SCSI packets to the application, producing significantly faster performance, lower network latency, lower access time, and lower CPU overhead. iSER can provide 6X faster performance than traditional TCP/IP based iSCSI. The iSER protocol unifies the software development efforts of both Ethernet and InfiniBand communities, and reduces the number of storage protocols a user must learn and maintain.

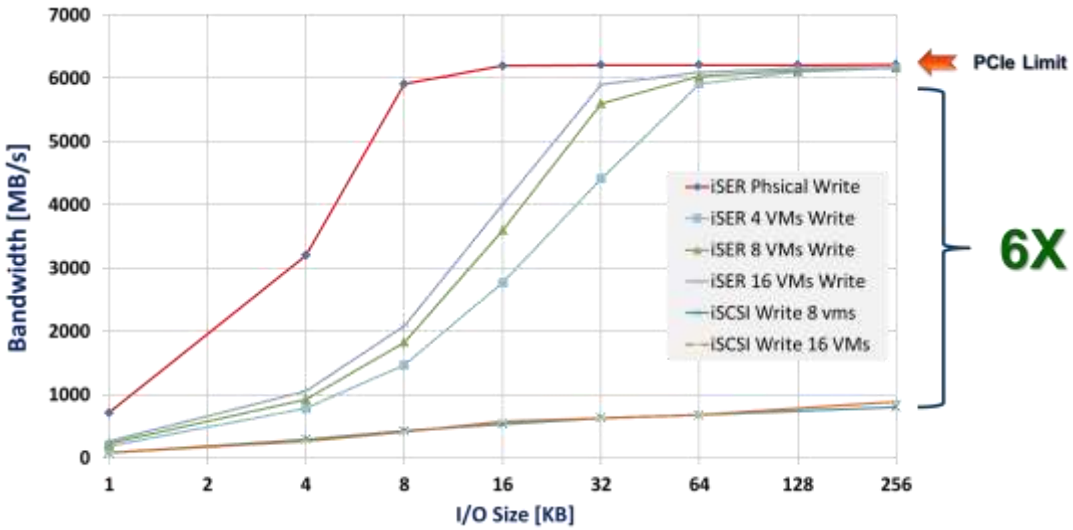
RDMA bypass allows the application data path to effectively skip to the front of the line. Data is provided directly to the application upon receipt without being subject to various delays due to CPU load-dependent software queues. This has the following three effects:

- The latency of transactions is incredibly reduced;
- Because there is no contention for resources, the latency is deterministic, which is essential for offering end-users a guaranteed SLA;
- Bypassing the OS using RDMA results in significant savings in CPU cycles. With a more efficient system in place, those saved CPU cycles can be used to accelerate application performance.

In Figure 4 it is clear that by performing hardware offload of the data transfers using the iSER protocol, the full capacity of the link is utilized to the maximum of the PCIe limit.

To summarize, network performance is a significant element in the overall delivery of data center services and benefits from high speed interconnects. Unfortunately, the high CPU overhead associated with traditional storage adapters prevents systems from taking full advantage of these high-speed interconnects. The iSER protocol uses RDMA to shift data movement tasks to the network adapter, thusly freeing up CPU cycles that would otherwise be consumed executing traditional TCP and iSCSI protocols. Hence, using RDMA-based fast interconnects significantly increases data center application performance levels.

Figure 4 - RDMA Acceleration



4 Networking

In this solution, we define the following node functions:

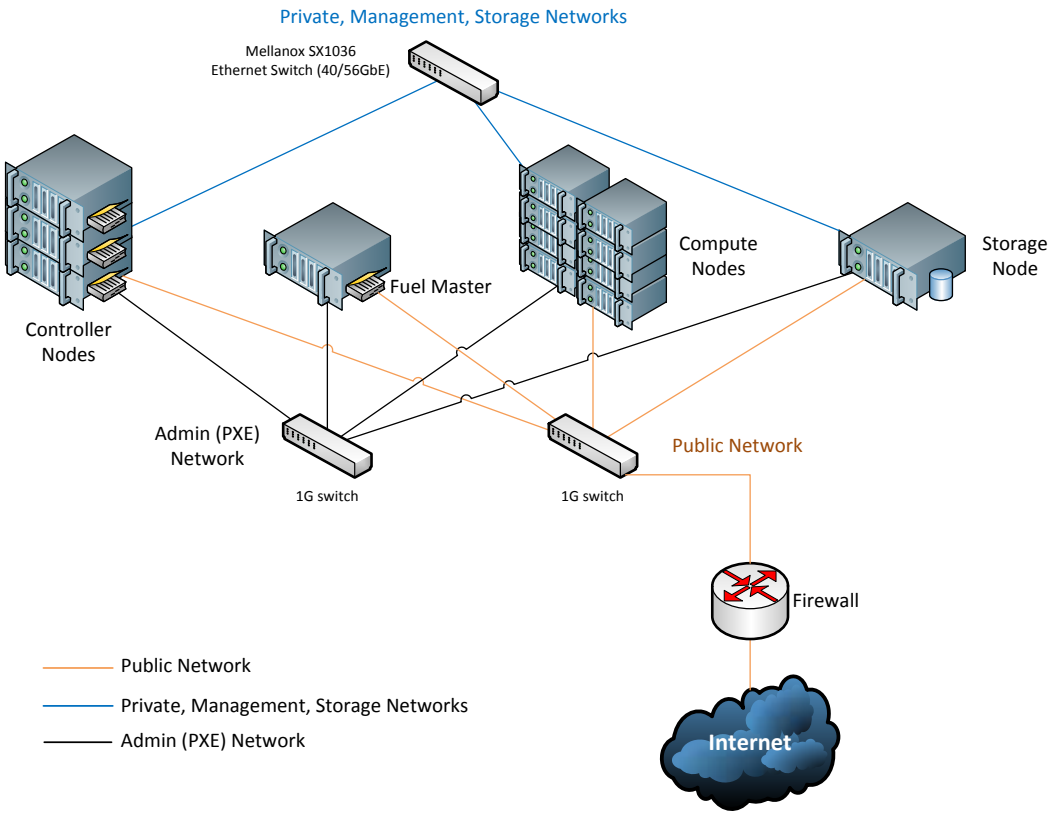
- Fuel node (master)
- Compute nodes
- Controllers (and network) node
- Storage node

The following five networks are required for this solution:

- Public network
- Admin (PXE) network
- Storage network
- Management network
- Private network

In this solution all nodes are connected to all five networks, besides the Fuel node that is connected to the Public and Admin (PXE) networks only. Although not all nodes may be required to connect to all networks, this is done by Fuel design. The five networks are implemented using three physical networks.

Figure 5 - Solution Networking

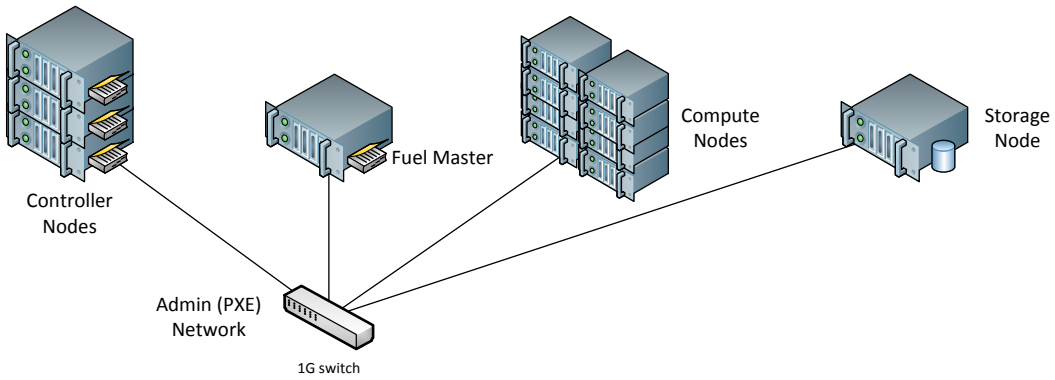


4.1 Network Types

4.1.1 Admin (PXE) Network

The Admin (PXE) network is used for the cloud servers PXE boot and OpenStack installation. It uses the 1GbE port on the servers.

Figure 6 - Admin (PXE) Network



4.1.2 Storage Network

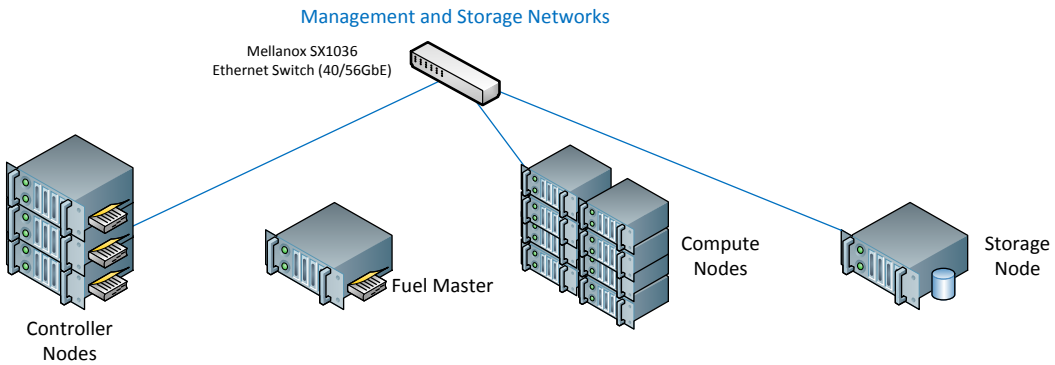
The storage network is used for tenant storage traffic. The storage network is connected via the SX1036 (40/56GbE switch).

The iSER protocol runs between the hypervisors and the storage node over the 40/56GbE storage network. The VLAN used for the storage network is configured by the Fuel UI.

4.1.3 Management Network

The Management network is an internal network. All OpenStack components communicate with each other using this network. It is connected with the SX1036 Ethernet switch (40/56GbE). The VLAN used for the management network is configured by the Fuel UI.

Figure 7 - Management and Storage Networks



4.1.4 Private Networks

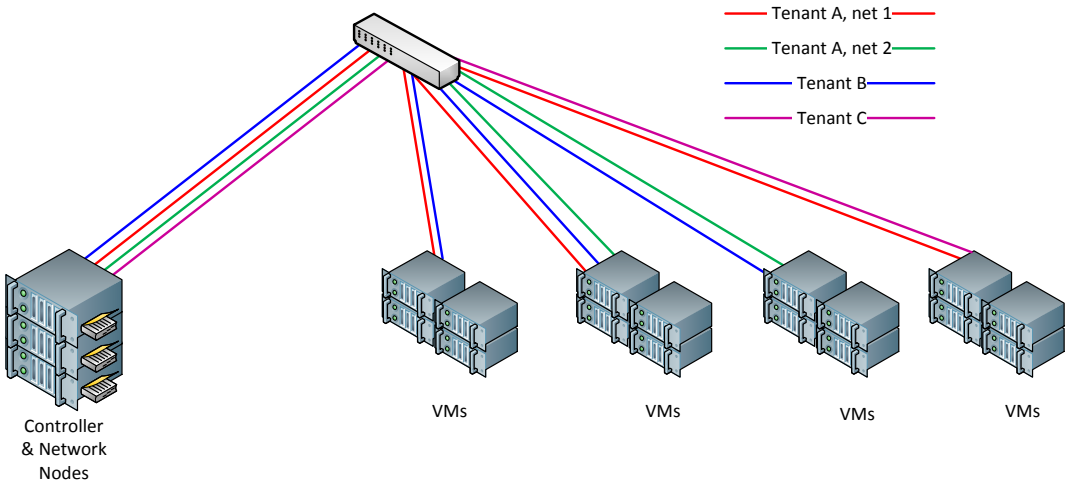
The private networks are used for communication among the tenant VMs. Each tenant may have several networks. If connectivity is required between networks owned by the same tenant, the traffic passes through the network node which is responsible for the routing. The VLAN used for the management network is configured by the Fuel UI.

Fuel 5.1/5.1.1 is based on OpenStack 'Icehouse' which only supports one network technology. This means that all the private networks in the OpenStack deployment should use Mellanox Neutron agent which is based on VLANs assigned to VFs.

The VLAN range used for private networks is configured by the Fuel UI.

Note: Allocate a number of VLANs according to the number of private networks to be used.

Figure 8 - Private Network



4.1.5 Public Network

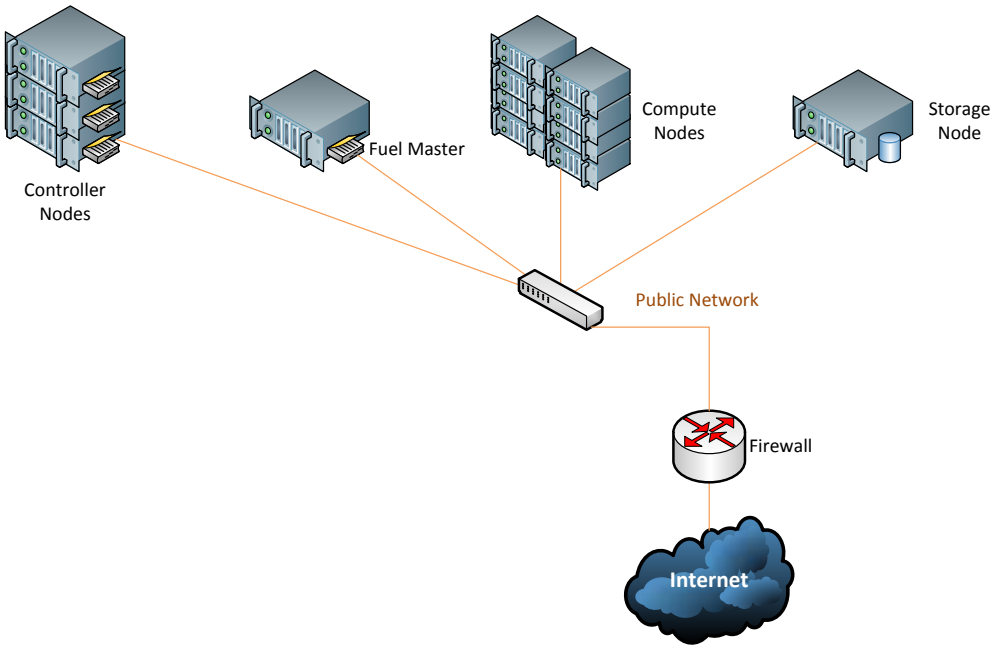
The public network enables external connectivity to all nodes (e.g. internet). The public network runs over the 1GbE ports of each server. This network is also used to access the different OpenStack APIs.

The public network range is split into two parts:

- Public range, which allows external connectivity to the compute hypervisors and all other hosts; and
- Floating IP range which enables VMs to communicate with the outside world and is a subset of addresses within the public network (via the controller node)

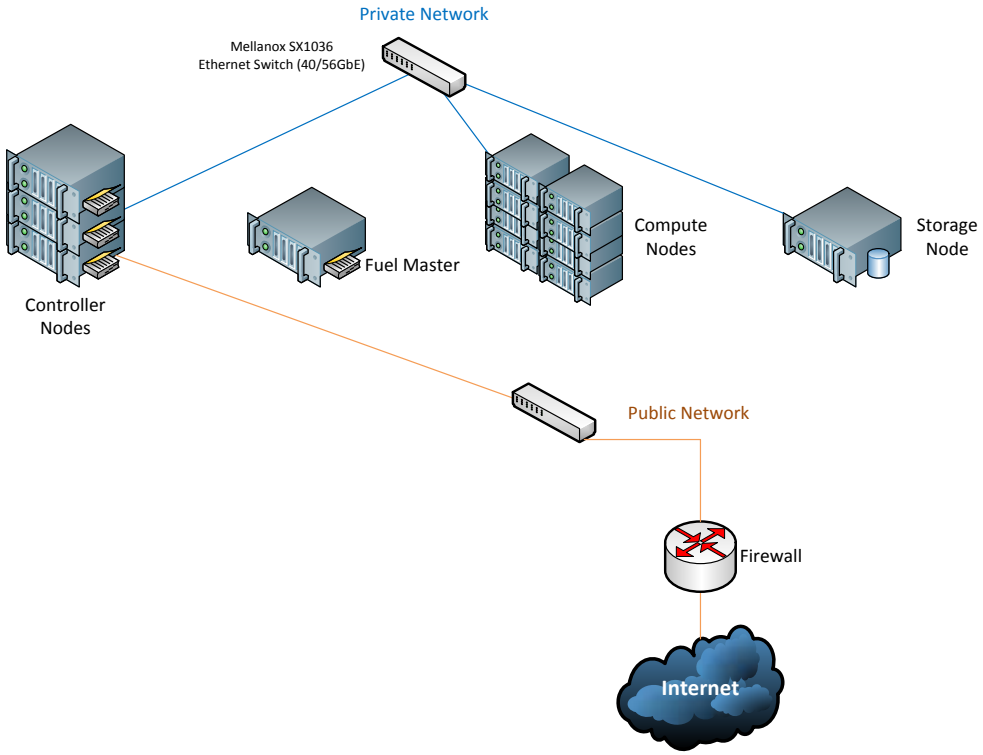
Another use for public access for hypervisors or node OS might be necessary (e.g. ssh access). However, the cloud owner must decide how to allow external connectivity to cloud servers, or cloud servers' access to the internet (either the PXE or public network may be used).

Figure 9 - Public Network



The floating IP uses the private network to pass traffic from the VMs to the network node and then the public network to pass traffic from the network node to the internet.

Figure 10 - Floating-IP Network



4.2 Physical Connectivity

The five networks discussed above can be connected via three ports in each server using 1GbE ports for admin (PXE) and public networks, and a single ConnectX®-3 Pro adapter port for private, storage, and management networks.

Figure 11 - Physical connectivity

4.3 Network Separation

VLANs should be configured on the switches to allow network separation. The VLANs to be configured on the switches should be aligned with the VLAN range configured in Fuel for the management, private, and storage networks.

4.4 Lossless Fabric (Flow-Control)

RDMA requires lossless fabric. To achieve that, flow control (global pause) should be enabled on all ports that may run RDMA and are connected to the SX1036 switch system. This is achieved by configuring global pause across all network hardware components.

5 Requirements

Mirantis Fuel defines the following node functions:

- Fuel node
- Compute nodes
- Controller nodes
- Storage nodes

5.1 Hardware Requirements

Table 1 lists the minimum requirements of hardware components:

Table 1 - Hardware Requirements

Component	Quantity	Description
Fuel node (master)	1	Intel server, 2 Gb Eth Ports, >4 cores CPU, >=8GB RAM, 0.5 TB SATA HDD
Compute nodes	2-32	Intel PCI-Ex Gen-3 server, at least one 8x PCI-Ex 3 slot, 2 Gb Eth Ports, >4 cores CPU x 2, >=128GB RAM RAM, 0.5 TB SATA HDD, SRIOV support in BIOS. ConnectX®-3 PRO EN or ConnectX®-3 PRO VPI Single Port network adapter. P/N (EN: MCX313A-BCCT or VPI: MCX353A-FCCT)
Controller nodes	3	Intel PCI-Ex Gen 3 server, at least one 8x PCI-Ex 3 slot, >2 x 1 Gb Eth Ports, >4 cores CPU x 2, >=32 GB RAM, 1 TB SAS HDD, SRIOV support in BIOS ConnectX®-3 PRO EN or ConnectX®-3 PRO VPI Single Port network adapter. P/N (EN: MCX313A-BCCT or VPI: MCX353A-FCCT)
Storage node	1	Intel PCI-Ex Gen-3 server, at least one 8x PCI-Ex 3 slot, 2 Gb Eth Ports, >4 cores CPU x 2, >=64GB RAM RAM, 0.5 TB SATA HDD ConnectX®-3 PRO EN or ConnectX®-3 PRO VPI Single Port network adapter. P/N (EN: MCX313A-BCCT or VPI: MCX353A-FCCT)
Storage, management, private switch	1	Mellanox SX1036 40/56GbE 36 ports
Public, admin (PXE) switch	1 or 2	1Gb switch (any switch)
56Gb/s cables	1 per server	FDR InfiniBand/56GbE copper cables up to 2m. P/N: MC2207130-XXX.
1Gb/s cables	2 per server	

Note: For an example of high-performance storage solution using RAID adapter, click here.

5.2 Operating Systems

All servers must be equipped with CentOS6.5 or Ubuntu 12.04.4 operating systems (Mirantis distro) via Fuel servers.

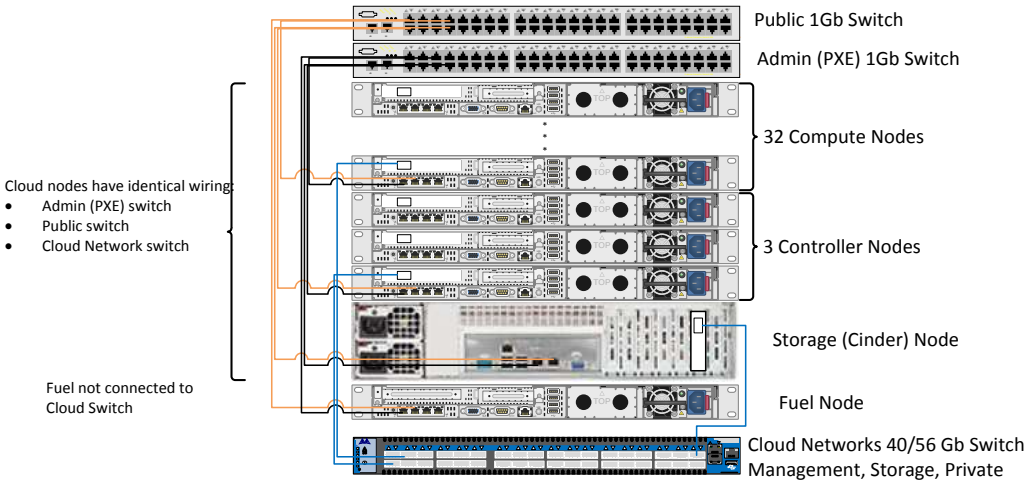
The VM should have the appropriate driver to support SR-IOV mode over Mellanox ConnectX-3.

6 Rack Configuration

6.1 32 Compute Node Setup with HA (3 Controller Nodes)

This section supplies a rack recommendation design for basic cloud setup up for 36 nodes (32 compute nodes).

Figure 12 - Rack Configuration



7 Installation and Configuration

For information about cloud installation and configuration click here.

For custom storage server installation and a configuration example click here.